DRAFT
Charge to EPA Science Advisory Board

Reproducibility Under the OMB Information Quality Guidelines

IntroductionBackground

In response to section 515 of Public Law 106-554, the White House Office of Management and Budget (OMB) issued *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Dissemination by Federal Agencies,* effective January 3, 2002. OMB directed each affected agency to develop their own guidelines to ensure and maximize the quality of information. EPA is now developing its own guidelines in preparation for the October 1, 2002 deadline. Reproducibility of data in influential government information is central to these guidelines, where OMB specifies that:

> "Agencies may identify, in consultation with the relevant scientific and technical communities, those particular types of data that can practically be subjected to a reproducibility[1] requirement, given ethical, feasibility, or confidentiality constraints..."

In this consultation with the SAB, EPA is particularly interested in ascertaining how the academic and scientific community addresses reproducibility in the conduct and publication of scientific research, viewed in the context of maintaining the quality of science. In addition to this primary focus, the SAB is requested to provide guidance on the related issue of approaches to applying rigorous robustness checks to analytic results when original data and models cannot be made available. Following this, EPA is interested in the extent to which such practices can be incorporated into scientific research and analyses developed, used, or supported by the Agency.

Background to the SAB Consultation

viewed in the context of maintaining the quality and objectivity of science. EPA is also interested in gaining insights into the extent to which such practices could be incorporated into EPA's information quality procedures.

Heightened federal attention to reproducibility traces to legal challenges regarding access to the primary data from major health studies of particulate air pollution, as well as isolated incidents of misconduct in other unrelated studies. These are concrete examples of a broader and more complex issue of reproducibility that spans several inter-related areas (e.g., quality management, science practice, legal requirements, information transparency). Multiple technical terms with shades of meaning also come into play (e.g., repeatable, replicable, reproducible), along with statutory, legal, and public policy determinations on what information should be made accessible to the public. In order to help reduce this complexity, this consultation with the SAB is focused on elucidating general practices adopted by the scientific community to validate or question published results – especially for research findings of major impact. Although recognizing that much of the political attention on reproducibility has emanated from legal challenges, this consultation should focus on general scientific practice, steering clear of judicial decisions and political controversies to the extent possible.

Currently, the transparency of information disseminated by the EPA is based on adherence to EPA Order 5360.1 A2, *Policy and Program Requirements for the Mandatory Agency-Wide Quality System,* and EPA Directive 2100, *Information Resources Management Manual, Chapter 10 _ Records Management*, July 1996. The Agency-wide Quality System incorporates by reference the American National Standard ANSI/ASQC E4-1994, and includes the necessary elements to plan, implement, document, and assess the effectiveness of QA/QC activities applied to environmental programs conducted by or for the EPA. The Order applies to the characterization of environmental or ecological systems and the health of human populations, direct measurements of environmental conditions or

[1] According to OMB's guidelines, "rReproducibility" means that the information is capable of being substantially reproduced, subject to an acceptable degree of imprecision. For information judged to have more (less) important impacts, the degree of imprecision that is tolerated is reduced (increased). With respect to analytic results, "capable" of being substantially reproduced" means that independent analysis of the original or supporting data using identical methods would generate similar analytic results, subject to an acceptable degree of imprecision or error.

releases, and the use of environmental data collected for other purposes or from other sources. Included under this latter category is the use of secondary data from the published literature, databases, or from models of environmental processes and conditions. Transparency of information is further supported by the development and use of standard operating procedures, standard analytical methods, and standard documentation and record keeping practices. Information and records management policies and procedures also play an important role in ensuring availability and access to the information necessary to achieve an appropriate degree of transparency.

Commensurate with the importance and legal mandate of the Agency action, and the impact of specific data, models or publications on such an action, Agency practice is to utilize best scientific judgment in weighing the quality of the data, the degree of transparency achieved in its presentation, and the extent to which the primary information and analyses can be relied upon in Agency decision-making. Major scientific and technical work products related to Agency decisions normally should be peer-reviewed. The elements in the Agency Quality System related to transparency (e.g., standard operating procedures, analytic methods, documentation, and record keeping procedures) apply to this information. In addition, under 40 C.F.R. §30.36(c), EPA may obtain, publish, or otherwise use data first produced under a grant or cooperative agreement. Under §30.36(d), in response to a FOIA request, EPA must request, and the recipient must provide, research data relating to published research findings produced under an award that were used by EPA in developing an agency action that has the force of law. The regulation narrowly defines "research data" to exclude certain materials, including confidential business information, personal privacy material, and similar information protected under law.

Whereas access to primary data is generally available from federally-performed or -funded research, the SAB's attention is particularly directed toward the large amount of important data and information in the published, peer-reviewed, literature, where funding may be from non-federal sources and quality standards are maintained by the scientific community. In earlier consultations with the National Academy of Sciences, increased emphasis was noted in peer reviewed journals (e.g., Science; Environmental Science and Technology) on providing electronic annexes of supplementary information to support study findings, without cluttering the primary paper with these details. These data are often in the form of summary tables and results, but not usually the primary data and certainly not to the level of laboratory records and chain of custody required in legal proceedings. Presentations were also made of various approaches for independently verifying study analyses in ways that maintain the intellectual property rights of the primary researchers. While the Agency recognizes that these are valuable options for enhancing transparency, they are not always available or feasible, and decisions on protecting public health and the environment may need to be based on less than ideal information. SAB reviewers are therefore encouraged to consider the spectrum of information sources and purposes to which information may be applied in their deliberations.

Background to EPA's draft "information quality guidelines"

EPA's draft "information quality guidelines" are consistent with the terms and concepts described in the OMB guidelines, including placing a special emphasis on "objectivity" and requiring, as a basic standard of quality, that disseminated information be objective in both presentation and substance. EPA's draft "information quality guidelines" rely largely on implementation of the previously described Agency policies and processes that are intended to ensure and maximize the quality and availability of information disseminated by the Agency to the public. With regard to reproducibility, EPA's draft guidelines state the following:

> "EPA recognizes that influential scientific, financial, or statistical information should be subject to a high degree of transparency about data and methods to facilitate the reproducibility of such information by qualified third parties, to an acceptable degree of imprecision. It is important that analytic results have a high degree of transparency regarding (1) the source of the data used, (2) the various assumptions employed, (3) the analytic methods applied, and (4) the statistical procedures employed. It is also important that the degree of rigor with which each of these factors is presented and discussed be scaled as appropriate, and that all factors be presented and discussed. In addition, if access to data and methods cannot occur due to compelling interests such as privacy, trade secrets, intellectual property, and other confidentiality protections, EPA should to the extent practicable, apply robustness checks to analytic results and document what checks were taken. Original and supporting data may not be subject to the high and specific degree of transparency required of analytic results; however, EPA should apply relevant Agency

policies and procedures to achieve reproducibility to the extent practicable, given ethical, feasibility, and confidentiality constraints."

With regard to influential original and supporting data, OMB suggests to agencies that if their guidelines do subject specific types of original or supporting data to a reproducibility requirement, then the associated guidelines should provide relevant definitions of reproducibility (e.g., standards for replication of laboratory data). With regard to influential analytic results, OMB suggests that agencies "generally require sufficient transparency about data and methods that an independent reanalysis could be undertaken" by a qualified party.

To better address the issue of reproducibility in the Agency's use of information from external sources (e.g., peer reviewed journals), EPA's draft guidelines express the intention to develop and publish factors that EPA could use in the future to assess the quality of voluntary submissions or information that the Agency gathers for its own use. Thus, this consultation with the SAB is also directed toward helping to inform the Agency's consideration of such factors that would enable future external providers of information to be aware of EPA's quality expectations.

Charge questions on reproducibility in the scientific community

1 What types of (original and supporting) data are generally expected to be reproducible in the scientific community?

< What does the scientific community consider to be a reasonable level of data provision/summary statistics or model parameters necessary to facilitate independent confirmation of results in the peer reviewed literature?
< How does the level of information expected vary with the importance of the study findings?
< What definitions or standards of reproducibility are used by the scientific community?
< How does the scientific community distinguish between reproducibility for (original and supporting) data vs. analytic results?
< How could these be applied by an agency to evaluating data and analytic results?

2 What is the role of peer review in addressing the reproducibility of data and results?

< To what extent can and do peer reviewers request and obtain primary data during their review of a submitted paper?
< What stipulations are placed on access to and use of these data?

3 Following publication, what constraints are there on the ability of researchers to make public their primary data, and what would be the impact on the researcher of making these data publicly available?

< What ethical, feasibility, and confidentiality constraints influence the ability to reproduce data?
< Are there other constraints that affect the provision of primary data?

4 What data omissions, discrepancies or other factors would lead to concern in the academic community about the veracity of a research finding?
< What types of robustness checks are applied in the academic community to assess published results?
< How would concerns about findings be addressed in an academic institution? the wider academic community?
< What options are available for gaining access to and reproducing primary data and analyses in scientific community?

5 Can the SAB suggest additional options for facilitating the reproducibility of information and/or access to primary or supporting data not otherwise available in the peer reviewed literature?